Ontology-based integration for bioinformatics

Vaida Jakonienė and Patrick Lambrix

Department of Computer and Information Science
Linköpings universitet, Linköping, Sweden
{vaija,patla}@ida.liu.se

Abstract

Information integration systems support researchers in bioinformatics to retrieve data from multiple biological data sources. In this paper we argue that the current approaches should be enhanced by ontological knowledge. We identify the different types of ontological knowledge that are available on the Web and propose an approach to use this knowledge to support integrated access to multiple biological data sources. We also show that current ontology-based integration approaches only cover parts of our approach.

1 Introduction

Researchers in bioinformatics often have to retrieve data from multiple biological data sources (DSs) to solve their research problems. Many such DSs are publicly available on the Web. For instance, 719 DSs are listed in the 2005 Database Issue of the Nucleic Acids Research [9] journal. As most DSs are developed and maintained independently, they are highly heterogeneous. They vary in the type of the stored data, the data format, and access methods. In addition, there is a terminology discrepancy at the data level and at the schema level, which even more complicates the data retrieval process. The user must decide which DSs to access and in which order, how to retrieve the data and how to combine the results - in short, the task of retrieving data requires a great deal of effort and expertise on the part of the user. Also, the users have to take into account that bioinformatics is a dynamic field where DS schemas change and new DSs are developed. Information integration systems (IISs) aim to alleviate these problems by providing a uniform (or even integrated) interface to underlying DSs. IISs may need to find DSs that are relevant to a user query, divide a query into smaller subqueries and, combine the retrieved results.

In addition to DSs, a large number of ontologies describing domain knowledge are publicly available in the area. OBO [10], an umbrella web address for ontologies covering the genomics and proteomics domains, lists 29 orthogonal ontologies. Some of the ontologies have reached the status of de facto standard and are used extensively to annotate DSs. Ontologies are also useful to support the integration of biological data and some of the current IISs incorporate ontology-related knowledge. However, this is still done in a limited way.

In this paper we argue that the available ontological knowledge on the Web should be used for the integration of biological data. First, we identify the different types of ontological knowledge that are publicly available on the Web in the field of bioinformatics (section 2). Then, we discuss how this knowledge can be used to support integrated access to multiple biological DSs (section 3). Further, we present an integration approach that combines the identified ontological knowledge with traditional information integration techniques (section 4). Finally, we show that current ontology-based integration approaches in bioinformatics cover parts of the suggested approach (section 5). Future work is given in section 6.

2 Ontological knowledge in bioinformatics

The publicly available ontological knowledge in bioinformatics includes: bio-ontologies, alignments between the ontologies, ontological annotations of DSs, and mappings between data values and ontological terms. We briefly describe each of these.

Bio-ontologies. There is a large variety of bio-ontologies. They differ in the type of biological knowledge they describe, their intended use, the adopted level of abstraction and the knowledge representation language. For instance, via OBO we can access a number of ontologies having different biological focus

Proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005

and that are developed for different purposes. GO ontologies describe biological process, molecular function and cellular components of genes and proteins in all organisms. The goal is to produce structured, precisely defined, common and dynamic ontologies that can be used for annotating gene products. MeSH is an ontology produced by the American National Library of Medicine and is used for indexing, cataloguing, and searching for biomedical and health-related information and documents. Anatomical Dictionary for the Adult Mouse (MA) is an anatomy ontology covering part of the lifespan of the laboratory mouse. The TAMBIS ontology [1] is an ontology covering a wide range of biological concepts and is used as a unified schema to support queries over multiple DSs in an IIS. With respect to the described knowledge abstraction the ontologies may range from high level ontologies that define general biological knowledge to ontologies that describe selected biological aspects. For instance, some general biological knowledge is covered in the TAMBIS ontology, like protein and nucleic acid are biomolecules, and motif is-component-of protein. On the other hand, the GO molecular function ontology defines the space of possible biological functions, like signal transducer activity and the more specific function receptor activity. Depending on the knowledge that is represented the ontologies can be classified from controlled vocabularies, taxonomies, thesauri, data models, and frame-based ontologies to knowledge-based ontologies [5]. These different types of ontologies can be represented in a spectrum of representation formalisms ranging from very informal to strictly formal. Many ontologies in bioinformatics started as controlled vocabularies, which are essentially list of terms (e.g. MeSH). Nowadays, a number of ontologies are augmented to support more advanced representation. For instance, GO and MA can be classified as thesauri, as they organize terms in a graph where the arcs in the graph represent a fixed set of relations. For instance, MA organizes anatomical structures spatially and functionally, using is-a and part-of relations (e.g. brain is-a head organ and it is part-of central nervous system). In addition, GO ontologies support the exact_synonym and narrow_synonym relations. The TAMBIS ontology can be classified as a knowledge base which is based on description logics.

Ontology alignments. The existing bioontologies may either contain overlapping information, provide different views on an area or may cover different areas. To combine different types of available knowledge, multiple ontologies may have to be used. To enable this, it is important to know relations, called alignments, between the terms in different ontologies. These alignments may describe equivalence, specialization or other relations between terms. For instance, the SOFG [11] resource on ontologies publishes alignments between SAEL, a high level ontology of cross-species anatomical terms, and ontologies of a single organism (e.g. MA). Currently, not so many inter-ontology alignments are available. In the near future we expect the increase of such knowledge as ontology alignment and merging tools are developed to support the identification of such alignments (e.g. [6]) For instance, given the terms auditory bone (MA) and ear ossicle (MeSH), and knowing that incus is a kind of auditory bone (MA), such a system would be able to identify that the given terms represent the same thing and to derive that incus is a kind of ear ossicle. The used matching techniques also enable identifying relations between completely different terms, e.g. that inner ear (MA) is a synonym to labyrinth (MeSH).

Annotations. To describe properties of biological objects in a uniform way, it becomes common in bioinformatics to annotate data entries in DSs with ontological terms. For instance, terms from GO molecular function ontology are used to describe gene and protein functions. Annotations can be stored as separate mapping rules, included in an ontology or stored in a DS entry. For instance, different DS annotations by GO terms can be found on the web pages of the GO Consortium. In addition to other relations, GO ontologies support the xref_analog relation that allows to link ontological terms to biological objects having the described properties.

Mappings between data values and ontological terms. In a similar way as whole data entries are related to ontological terms, allowed values for certain data properties can be indexed based on ontology terms. For instance, keywords used to describe data entries in UniProt, a DS of protein sequences and related data, are mapped to terms in GO ontologies. Similarly as for ontology alignments, different techniques could be used to support the identification of matching terms.

3 Integration support through the available knowledge

Some of the important steps in querying over multiple DSs are user query formulation, DS selection together with the query rewriting into subqueries over the selected DSs, and identification of relevant data items on which results from different subqueries can be joined. In this section we describe how the ontological knowledge identified in section 2 can be used to support these steps.

Query formulation. Ontologies can be used for guiding users through query formulation. An IIS can provide an ontology as a query formulation interface or can support inclusion of ontology terms into a query. High level ontologies enable the selection of relevant types of biological knowledge, while specialized ontologies (e.g. GO molecular function ontology) can be useful for the precise specification of properties for data items of interest. Different ontolo-

gies support querying from different points of view, e.g. query for genes involved in a biological process or genes expressed in a particular cellular component. As the user query may cover different types of biological knowledge that is spread over a number of ontologies, mapping rules between ontologies are important. This enables ontological reasoning over different domains. For instance, such rules would allow reasoning based on relationships between protein function and diseases. An important use of ontologies is for query expansion. This leads to better query results. When queries are expanded using terms equivalent to the query terms, the terminology discrepancy problem is reduced. When generalization-specialization relationships are used, more relevant results are retrieved. For instance, knowing that receptor is-a signal transducer, a query asking for specific signal transducers can be expanded to retrieve receptors having the same properties. Also, checking query validity can be performed with respect to the domain knowledge.

Data source selection and query rewriting. The ontological knowledge is important for describing DSs uniformly from the domain perspective. When user queries include ontological terms, such DS descriptions provide support for DS selection and the user query rewriting into subqueries over DSs. Terms from high level ontologies can specify types of biological data stored in DSs such as, for instance, that sequences stored in UniProt represent protein sequences. At the same time relations between data items in a DS could be derived by the available relations between ontological terms (e.g. domain is a part-of sequence). Specialized ontologies could be used to specify the range of possible values for a data type (e.g. which organisms are covered in a DS). Also, ontological terms can be used to refine the description of the content of a DS. Often, not all data is stored explicitly in a biological DS. For instance, in a DS capturing mouse related data, mouse will not be mentioned explicitly in the data entries. In addition, the knowledge about existing ontological annotations of DSs and mapping rules between DS and ontology terms should be used to specify the DS schema. Ontological annotations are useful for fast and focused searches on a certain type of data (e.g. search UniProt on GO terms describing protein function). The annotations directly lead to relevant data entries in DSs. The disadvantage is that such searches only enable the retrieval of annotated data. Exhaustive and complex searches should be complemented by other data querying techniques. Mapping rules between DS and ontology terms provide a basis for translating query constraints expressed over ontologies to DS specific terms.

Data integration. When results for the subqueries are retrieved from different DSs, the next step is to identify which data entries can be joined. A straightforward approach is to require equality between the joined data items. As there is no agreed terminology and there are no unique identifiers for terms in bioinformatics, often this approach is not suitable. The joined data items may have different but synonymous data values, they can be described at different levels of granularity or use different lexical variations. For instance, a gene can be referred to by different identifiers, like insulin promoter factor, a gene activating transcription of insulin genes, is named IPF1, IDX-1, STF-1, PDX-1, PDX1 and MODY4. For an organism its scientific or common name can be used, e.g. mus or mice. In some DSs the type of organism could be specified more precisely, e.g mus famulus. Also, B Cell Leukemia can be written as Leukemia, B Cell. Ontological knowledge provides a range of possibilities on how to handle these issues. Joins could be performed on the basis of ontological terms. Mapping rules between DS and ontology terms could be used to translate values into a uniform representation. Also, data can be joined on the available ontological annotations. Further, ontological knowledge about synonyms can be used to locate alternative data item representations. To cover different granularity of data items, is-a relationships in ontologies should be explored. Mapping rules between ontologies should be used to combine data items retrieved from different domains.

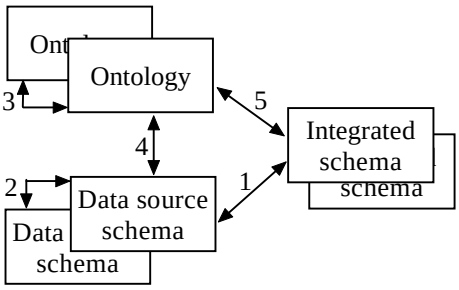

Figure 1: Types of supported knowledge and mapping rules between them.

4 Ontology-based data source integration

This section presents an approach for ontology-based support for access to multiple biological DSs. In this paper we focus on how knowledge should be set up to support query processing in the system. The primary goal is to use the publicly available ontological knowledge in bioinformatics (section 2) to better support query processing (section 3).

Figure 1 represents the main types of knowledge used in our approach as well as the existing mapping rules (MR) between them. The main types of knowledge in traditional IISs are DS schemas (DSSs) and integrated schemas (ISs), where an IS combines the relevant domain knowledge with data structures contained in the integrated DSs. Such systems enable DS integration through e.g. global-as-view or local-

as-view rules (MR 1). Also, existing cross-references between data entries at different DSs are considered to enable joins of retrieved data (MR 2). As we have shown in the previous section, the ontological knowledge in bioinformatics may provide extra information and support querying over multiple DSs. Based on this, we suggest that in addition to the traditionally used knowledge, domain ontologies, ontological alignments (MR 3) and ontology-based DS descriptions (MR 4) should be used. Here, DS descriptions cover ontological annotations, mappings between data values and ontological terms together with ontology-based DSS descriptions. To reuse the existing ontological knowledge and to uniformly specify the integrated data, ontology based IS descriptions, similarly to DS descriptions, should be maintained (MR 5). To provide a better focus on certain types of research questions, several ISs may be supported.

The use of the available and autonomously maintained domain knowledge may lead to faster development and easier maintenance of IISs in a dynamic environment. For instance, the ontological alignments may lead to smaller ISs needed for integration. Also, as ontologies are maintained autonomously, the chance that we need to modify ISs when the biological knowledge is updated, decreases. Further, for users it may be easier to comprehend an IIS as in many cases they are familiar with the available ontologies. In addition to query answering scenarios found by traditional IISs, the ontological knowledge may suggest other ways to solve the task. In some cases one type of scenario may be preferable over another while in other cases they need to be used in combination to get a complete set of answers. To handle this, strategies for managing alternative query answering possibilities have to be developed. Further, it is important to choose or develop suitable representation and reasoning mechanisms for the different types of knowledge and mapping rules.

5 Comparison to Current Approaches

There are several IISs for integrating biological DSs (for an overview see [2]). Some of these IISs use ontology-based technologies to support querying (e.g. BACIIS [8], KIND [7], SEMEDA [3] and TAMBIS [1]). A common feature is that the ISs used in these systems are seen as ontologies. In contrast, we expect ontologies to be agreed upon and shared by many users [5]. As in our approach, the ISs include domain knowledge and information on data structures at the DSs. All the systems use the maintained ontology to describe the content of DSs (MR 1). Though it is not explicitly stated, cross-references between DSs are probably used to join the retrieved data items (MR 2). KIND uses two ontologies describing static and process knowledge, respectively. The ontologies combine domain knowledge from neuroanatomy and neurophysiology (MR 3). In SEMEDA controlled vocabularies can be used to specify semantics of data type values, which covers part of MR 4 in our framework. Also, DS content descriptions can be refined with IS terms. Ontological annotations and mappings between ontology terms are not taken into account in any of the systems.

6 Future Work

As a next step in our project we plan to implement our ideas into the BioTRIFU system [4]. We will define a number of test cases and scenarios. Further, we will investigate formalisms to represent DS descriptions and the different types of mapping rules as well as relevant reasoning mechanisms.

Acknowledgements This research work was funded by the Swedish national graduate school in computer science and CENIIT. We also acknowledge the support of the EU Network of Excellence REWERSE (Sixth Framework Programme project 506779).

References

[1] Goble CA, Stevens R, Ng G, Bechhofer S, Paton N, Baker P, Peim M, Brass A (2001) Transparent access to multiple bioinformatics information sources. *IBM Systems Journal* 40(2).

[2] Jakonienė V (2005) *A Study in Integrating Multiple Biological Data Sources.* Licentiate thesis No 1149, Linköpings universitet, Sweden.

[3] Köhler J, Philippi S, Lange M (2003) SEMEDA: ontology based semantic integration of biological databases. *Bioinformatics* 19(18):2420-2427.

[4] Lambrix P, Jakonienė V (2003) Towards Transparent Access to Multiple Biological Databanks. *Proceedings of the Asia-Pacific Bioinformatics Conference*, pp 53-60.

[5] Lambrix P (2004) Ontologies in Bioinformatics and Systems Biology. Chapter 8 in Dubitzky W, Azuaje F (eds) *Artificial Intelligence Methods and Tools for Systems Biology*, pp 129-146, Springer.

[6] Lambrix P, Tan H (2005) Merging DAML+OIL Ontologies. Barzdins J, Caplinskas A (eds) *Databases and Information Systems*, pp 249-258, IOS Press.

[7] Ludäscher B, Gupta A, Martone ME (2003) A Model-Based Mediator System for Scientific Data Management. Chapter 12 in Lacroix Z, Critchlow T (eds) *Bioinformatics: Managing Scientific Data*, pp 335-370, Morgan Kaufmann Publishers.

[8] Miled ZB, Webster YW, Liu Y, Li N (2003) An Ontology for Semantic Integration of Life Science Web Databases. *International Journal of Cooperative Information Systems* 12(2):275-294.

[9] Nucleic Acids Research. http://nar.oupjournals.org

[10] Open Biomedical Ontologies. http://obo.sourceforge.net/

[11] Standards and Ontologies for Functional Genomics. http://www.sofg.org/